Comparative analysis of classical machine learning algorithms for binary classification of drinking water portability

Vasily A. Fartukov[1*], *Marina I.* Zborovskaya[1], and *Kristina S.* Semenova[1]

[1] Russian State Agrarian University – Moscow Timiryazev Agricultural Academy, Moscow, Russian Federation

Abstract. Today, drinking-quality water is becoming a scarce resource, where water quality serves as an indicator of anthropogenic impact on the environment. This study presents a solution to the problem of binary classification of drinking water based on hydrochemical parameters, utilizing a comparative analysis of classical machine learning algorithms. The relevance of this research stems from the need for rapid, automated methods for the preliminary assessment of water quality within the framework of decision support systems for water treatment and monitoring. The objective of the study was to identify the most balanced algorithm for classifying water as either suitable or unsuitable for drinking, based on standard physicochemical parameters. Evaluating water against these parameters aids in assessing the state of the aquatic environment and its ecological safety. To address this classification task using machine learning algorithms, a complete data analysis pipeline was implemented. Three algorithms were subjected to detailed examination: Decision Tree, Random Forest, and Gradient Boosting. Automated modeling was employed using the PyCaret library. All models were evaluated on a dedicated test dataset using the following metrics: Accuracy, Precision, Recall, F1-score, and AUC-ROC. The most balanced results were obtained with the Random Forest model, which achieved an Accuracy of 0.61 and an F1-score of 0.48. The Gradient Boosting model demonstrated comparable overall accuracy, while the Decision Tree model yielded the highest Precision score, reaching 0.71. A feature importance analysis for the Random Forest model revealed that pH is the most significant predictor, a finding consistent with fundamental principles of drinking water quality assessment. These results establish a foundation for the preliminary identification of potentially substandard water samples and for supporting water quality monitoring efforts.

1 Introduction

Ensuring access to safe drinking water is one of the key priorities in the context of public health protection and sustainable development [1]. Drinking water quality is regulated by

* Corresponding author: vasfar@mail.ru

strict standards that define maximum permissible concentrations for a wide range of physicochemical and microbiological parameters [2]. Traditional laboratory-based control methods provide high accuracy but require significant time and financial resources, which limits their applicability for large-scale and real-time monitoring [3].

The rapid development of digital technologies and the availability of open datasets aggregating hydrochemical analysis results have created favorable conditions for applying data-driven methods to preliminary and rapid assessment tasks [4]. Classical machine learning algorithms, characterized by a high degree of interpretability, relatively low computational requirements, and proven effectiveness on structured data, represent a promising tool for building automated water quality screening systems [5,6].

The problem of drinking water potability assessment can be formalized as a binary classification task, where the input consists of a vector of standard hydrochemical characteristics and the output corresponds to assigning a water sample to either the "potable" (1) or "non-potable" (0) class. For such tasks, decision-tree-based algorithms and their ensembles (Random Forest and Gradient Boosting) are widely used due to their robustness and their ability to estimate feature importance [7].

The aim of this study is to perform a comparative analysis of the performance and balance of classical machine learning algorithms for the binary classification of drinking water potability based on hydrochemical parameters.

To achieve this aim, the following objectives were defined:

1. Conduct an exploratory analysis of the initial dataset, and perform comprehensive, methodologically sound data preprocessing, including a stratified split into training and test sets. Train, tune, and validate Decision Tree (DT), Random Forest (RF), and Gradient Boosting (GB) models, and apply the PyCaret framework for additional automated modeling.

2. To perform a comparative evaluation of all models using a unified set of metrics on an isolated test set, identify the most informative features, and analyse error patterns.

Ensuring water quality is an important objective for environmental safety and sustainable water resource management. The method, thanks to the early identification of problematic samples, allows for more efficient allocation of laboratory and management resources.

2 Materials and Methods

2.1 Dataset Description

The study employed a publicly available anonymized dataset for drinking water potability classification (Water Potability Dataset) [8]. The dataset contains 3,276 records, each described by ten attributes: nine independent physicochemical features and one binary target variable, Potability (Table 1).

Table 1. Description of features in the dataset.

No.	Feature	Description	Unit	Data type
1	pH	Hydrogen ion concentration	–	Continuous
2	Hardness	Total hardness	mg-eq/L	Continuous
3	Solids	Total dissolved solids (TDS)	ppm	Continuous
4	Chloramines	Chloramine concentration	ppm	Continuous
5	Sulfate	Sulfate ion concentration	mg/L	Continuous
6	Conductivity	Electrical conductivity	µS/cm	Continuous
7	Organic_carbon	Organic carbon content	ppm	Continuous
8	Trihalomethanes	Trihalomethane concentration	ppb	Continuous

9	Turbidity	Turbidity	NTU	Continuous
10	Potability	Target variable (0 – non-potable, 1 – potable)	–	Binary

The target variable Potability reflects overall drinking water suitability based on compliance with key regulatory thresholds typical of international recommendations [2].

2.2 Exploratory Data Analysis (EDA)

The analysis was conducted using Python libraries: pandas for tabular data processing, numpy for numerical computations, and matplotlib and seaborn for visualization [9]. Initial data inspection revealed missing values in three features: pH (14.99%), Sulphate (23.84%), and Trihalomethanes (4.95%).

Basic descriptive statistics (mean, standard deviation, minimum, maximum, and quartiles) were calculated for all continuous variables. The proximity of mean and median values for most features indicates the absence of pronounced skewness. Boxplots were used to assess data dispersion and detect outliers; no critical anomalies requiring removal were identified.

Analysis of the target variable distribution revealed a significant class imbalance: 1,998 samples (61%) belong to the “non-potable” class (0), while 1,278 samples (39%) belong to the “potable” class (1). This factor was considered when selecting evaluation metrics and training models. The baseline metric corresponds to predicting the majority class, yielding an Accuracy of 0.61.

A Pearson correlation matrix was constructed to assess linear relationships between independent variables. The maximum correlation coefficient did not exceed 0.20, indicating the absence of strong linear dependencies and allowing all features to be used in modeling without a high risk of multicollinearity [10].

2.3 Data Preprocessing

2.3.1 Missing value handling

For features containing missing values, mean imputation was applied using values calculated exclusively from the training dataset. This strategy was chosen due to the approximately symmetric feature distributions.

2.3.2 Train–test split

After data cleaning, the dataset was divided into a feature matrix (X) and a target vector (y). Using the train_test_split function from scikit-learn [11], the data were split into training (80%) and test (20%) sets with random_state = 42 and stratification by the target variable.

2.3.3 Feature scaling

Z-score normalization (StandardScaler) was applied to bring all numerical features to a common scale and improve algorithm convergence. Scaling parameters were fitted only on the training set and then applied to the test set to prevent data leakage.

2.4 Machine Learning Algorithms and Evaluation Methodology

The following classical algorithms were implemented for the binary classification task.

2.4.1 Decision Tree Classifier (DT)

A simple and interpretable algorithm. To prevent overfitting, hyperparameters were tuned using GridSearchCV with 5-fold cross-validation.
Optimal parameters: max_depth = 5, min_samples_split = 10, criterion = 'gini'.

2.4.2 Random Forest Classifier (RF)

Random Forest Classifier (RF): An ensemble method that is resistant to retraining and allows you to assess the importance of signs. Setting of hyperparameters is carried out similarly. Лучшие параметры: n_estimators=200, max_depth=10, min_samples_leaf=4, class_weight='balanced'.

2.4.3 Gradient Boosting Classifier (GB)

Powerful ensemble method. Used HistGradientBoostingClassifier implementation with customization. The best parameters are max _ iter = 300, learning _ rate = 0.05, max _ depth = 5, l2 _ regularization = 1.0.

2.4.4 Automated modeling with PyCaret

For comparison, the PyCaret framework (version 3.0) was used [12], which automates the processes of preprocessing, model selection and setting up hyperparameters. The SMOTE technique is activated in the settings to eliminate class imbalances. The Voting Classifier is automatically created and optimized based on the Extra Trees and Random Forest algorithms.

2.4.5 2.4.5 Training and validation scheme

- Pre-setting of hyperparameters for DT, RF, GB was performed using GridSearchCV with 5x cross-validation on the training set.
- The final training of models with the best parameters was carried out on the entire training set.
- Final evaluation and comparison of all models (including the PyCaret model) was performed only on an isolated test sample not involved in tuning.

2.4.6 Evaluation of metrics

The following standard classification metrics were used [13]:
Accuracy: (TP+TN) / (P+N).
Precision: TP / (TP+FP).
Recall: TP / (TP+FN).
F1-score
AUC-ROC
Confusion Matrix (FP, FN).

3 Results

Comparative classification results.

The results of all four approaches on an isolated test sample (n = 655) are presented in Table 2.

Table 2. Comparative results of algorithms on the test sample.

Algorithm	Accuracy	Precision	Recall	F1-score	AUC-ROC
Decision Tree (DT)	0.58	0.71	0.28	0.40	0.57
Random Forest (RF)	0.61	0.62	0.48	0.48	0.63
Gradient Boosting (GB)	0.60	0.60	0.45	0.46	0.61
PyCaret Ensemble	0.61	0.58	0.52	0.47	0.62

For clarity of comparison, Figure 1 shows ROC curves for all models.

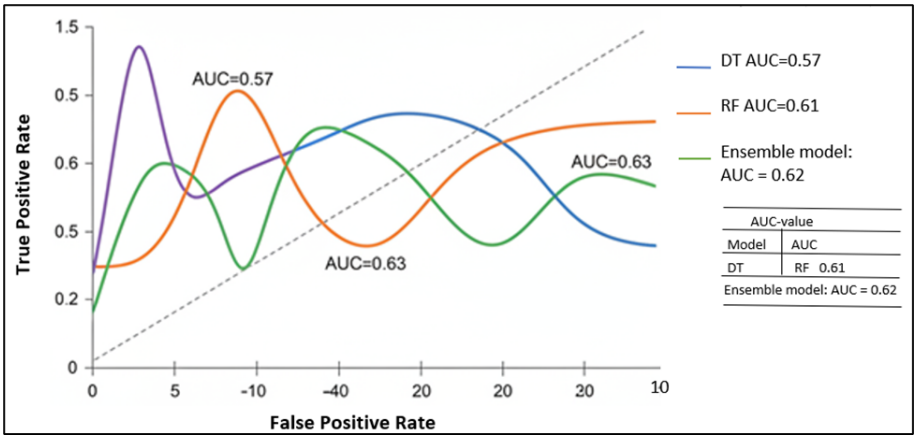

Fig. 1. ROC curves of the compared models on the test dataset.

3.1 The analysis of Table 2 supports the following conclusions

1. Random forest (RF) demonstrated the most balanced performance, showing best or near best values for all metrics except Precision. It reached maximum Accuracy (0.61), Recall (0.48), F1-score (0.48) and AUC-ROC (0.63). The high Recall compared to other models means that RF better identifies samples of usable water (minimizes false negative errors)..

2. Decision Tree (DT) achieved the highest Precision (0.71), indicating the lowest proportion of false-positive predictions among all models. However, its extremely low Recall (0.28) suggests that the model is overly conservative, classifying most samples as non-potable, which results in a low F1-score (0.40).

3. Gradient Boosting (GB) and the PyCaret ensemble showed performance comparable to Random Forest in terms of overall metrics (Accuracy ≈ 0.60–0.61, F1 ≈ 0.46–0.47). The PyCaret model, trained with class imbalance handling (SMOTE), achieved the most balanced Precision–Recall trade-off among all models (0.58 and 0.52, respectively).

3.2 Feature Importance Analysis

Feature importance analysis based on the Random Forest model identified the following ranking:

1. pH (0.24)
2. Solids (0.15)
3. Hardness (0.13)
4. Sulfate (0.10)
5. Organic carbon (0.09)
6. Trihalomethanes (0.08)
7. Conductivity (0.07)
8. Turbidity (0.07)
9. Chloramines (0.07)

The dominant role of pH has a clear scientific basis, as it strongly influences corrosion processes, disinfection efficiency, chemical equilibria, and sensory water properties, and is emphasized in all major drinking water guidelines, including those of the WHO [2].

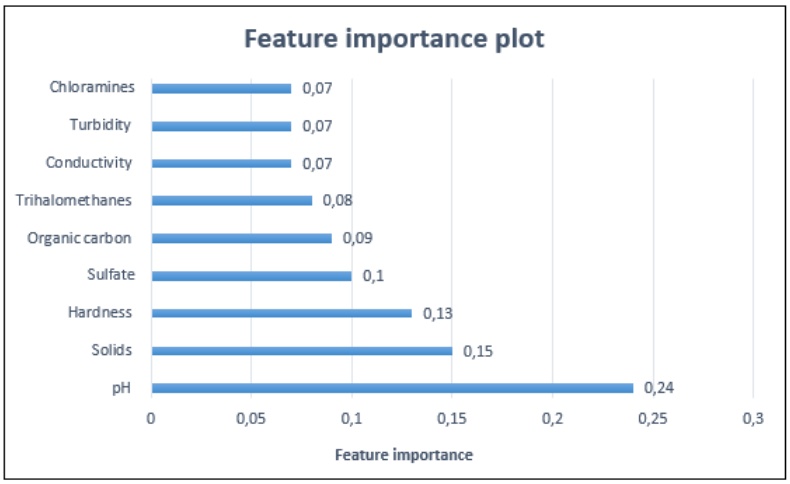

Fig. 2. Feature importance plot for the Random Forest (RF) model.

3.3 The leading role of pH has clear scientific justification

The hydrogen ion concentration is a fundamental parameter influencing water corrosivity, disinfection efficiency, chemical speciation, and organoleptic properties, which is reflected in all major regulatory frameworks, including the WHO guidelines [2]. The high importance of total dissolved solids (Solids) and water hardness is also consistent with practical experience, as these parameters directly affect the consumer-related properties of drinking water.

3.4 Error analysis

The confusion matrices for the two most effective models (RF and the PyCaret ensemble) are presented in Table 3.

Table 3. Confusion matrices for the RF and PyCaret models on the test dataset (absolute values).

Model	True class	Predicted: 0 (Non-potable)	Predicted: 1 (Potable)
Random Forest (RF)	0 (Non-potable)	TN = 235	FP = 56
	1 (Potable)	FN = 63	TP = 58
PyCaret Ensemble	0 (Non-potable)	TN = 220	FP = 71
	1 (Potable)	FN = 63	TP = 58

3.5 From a practical perspective, two types of classification errors can be distinguished

False Positive (FP): non-potable water is incorrectly classified as potable. This is the most critical type of error from a public health perspective. The Random Forest model produces 56 such errors out of 291 truly non-potable samples (19.2%), which is lower than that of the PyCaret ensemble (24.4%).

False Negative (FN): potable water is incorrectly rejected. This error is less critical in terms of health risks but economically significant, as it leads to unnecessary additional treatment costs. The Random Forest model produces 63 such errors out of 121 truly potable samples (52.1%), whereas the SMOTE-balanced PyCaret ensemble demonstrates a better result (47.9%).

3.6 The balance between FP and FN errors is determined by the classification threshold

In this study, the standard threshold of 0.5 was used. In practice, this threshold can be adjusted by increasing it to reduce FP errors (enhancing safety) or decreasing it to reduce FN errors (improving economic efficiency), depending on the priorities of a specific monitoring system

4 Discussion

The conducted comparative analysis revealed the strengths and limitations of classical machine learning algorithms for addressing the specific applied task of binary classification of drinking water potability.

1. Interpretability vs. performance balance: The Balance Between Interpretability and Performance: A simple Decision Tree (DT) model demonstrated the highest Precision and is suitable for developing validation mechanisms in scenarios were minimizing false positives (FP) is critical. However, its low Recall practically limits the model's applicability in automated systems.. The Random Forest (RF) and Gradient Boosting (GB) models, as more complex ensemble methods, provided substantially more balanced predictive performance (with F1-scores higher by approximately 17–20%), sacrificing some interpretability at the level of individual decisions while retaining the ability to analyze feature importance at the overall model level.

2. Impact of class imbalance and mitigation methods: The pronounced class imbalance in the dataset (61%/39%) had a substantial effect on the evaluation metrics, particularly on Recall for the minority class (“potable”). The DT, RF, and GB models trained without explicit imbalance handling exhibited a tendency to bias predictions toward the majority class. The application of the SMOTE technique within the PyCaret framework enabled the ensemble model to achieve the best Precision–Recall balance for the minority class, confirming the necessity of applying appropriate imbalance-handling methods when dealing with imbalanced data in applied tasks.

The obtained results for Random Forest (Accuracy = 0.61 and F1-score = 0.48) are consistent with findings from other studies using the same dataset, where the best-performing models rarely exceed an Accuracy of 0.65 [14, 15]. It should be noted that the achieved accuracy matches the majority-class baseline (Accuracy = 0.61), which highlights the difficulty of the task; however, the Random Forest model provides a more informative and balanced classification than the baseline strategy, as evidenced by its improved Recall, F1-score, and AUC-ROC.

3. Automated machine learning (AutoML) as a tool: The PyCaret framework enabled the rapid development of a model that is competitive with carefully hand-tuned algorithms. Its primary advantage lies in the automation of the entire modeling process, including the handling of class imbalance. The resulting ensemble is constructed using Extra Trees and Random Forest algorithms, which ensures robustness. A significant limitation is the "black box" nature of the configuration process, as there are a number of applications that require a sufficient degree of transparency and control.

4. Practical applicability and model selection: For deployment in a preliminary drinking water quality screening system, the Random Forest method represents the optimal compromise. It combines: sufficiently high and well-balanced predictive performance (the highest F1-score); robustness to overfitting; the ability to analyze feature importance, which is critically important for explaining results to domain experts and for validating the model from a subject-matter perspective (including the confirmed leading role of pH); acceptable computational efficiency for operational use.

In cases where safety is the primary concern (i.e., minimizing false positive errors), a Decision Tree model with high Precision or a Random Forest model with an increased classification threshold may be considered.

5. Limitations and reasons for moderate performance: It should be noted that the absolute values of the evaluation metrics (Accuracy $\approx$ 0.61, F1 $\approx$ 0.48) indicate that classification based on nine physicochemical parameters is a non-trivial task. This can be attributed to several factors: (a) drinking water standards are inherently complex and threshold-based rather than linear; (b) some parameters critical for specific contamination scenarios (e.g., microbiological indicators or heavy metal concentrations) may be absent from the dataset; and (c) the original data may contain noise or incomplete labels. Such results are typical for problems of this type, where decisions are made based on a limited set of parameters without incorporating contextual information (e.g., water source, seasonality, or geographical factors) [16].

5 Conclusion

The study included a full cycle of work on constructing, fine-tuning, and comparative analysis of classical machine learning algorithms for the task of binary classification of water suitability for drinking.

5.1 Key results and conclusions

A comprehensive analysis methodology was developed and implemented, including missing value handling, feature-wise scaling, stratified data splitting, and a unified training and evaluation scheme using an isolated test dataset.

The fundamental applicability of classical machine learning algorithms to this task was confirmed. The Random Forest method demonstrated the most balanced performance (F1-score = 0.48, AUC-ROC = 0.63) and is recommended as an optimal baseline algorithm for building screening systems that balance accuracy, robustness, and interpretability.

The Decision Tree model achieved the highest Precision (Precision = 0.71), making it a suitable candidate for developing simple decision rules in applications where minimizing false positive errors is critical.

The use of an automated approach (PyCaret) with the application of SMOTE enabled effective handling of class imbalance and resulted in a model with the best Precision–Recall balance for the minority class, confirming the importance of explicitly accounting for this factor.

A key scientifically grounded result is the identification of the hydrogen ion concentration (pH) as the most influential feature in the Random Forest model, which fully aligns fundamental principles of drinking water quality assessment and confirms the domain adequacy of the model.

5.2 Future research directions include the following

Integration of additional data sources, including microbiological indicators and spectroscopic data, to improve the completeness of system characterization;

Development of multiclass classification models (e.g., quality categories such as excellent, good, acceptable, and non-potable) to enable more detailed assessment;

Experimentation with neural network architectures (e.g., simple fully connected networks) to compare their performance with classical methods for this task;

Development of a prototype web application or API service based on the best-performing model (Random Forest) for rapid preliminary assessment of drinking water quality using input parameters, which has direct practical relevance for specialists in water treatment and environmental monitoring.

The conducted study confirms the practical significance of classical, interpretable machine learning algorithms as effective tools for developing decision-support and auxiliary analytical systems in the critically important field of drinking water quality control. Thus, classical machine learning methods can serve as a relevant tool for preliminary water quality screening, supporting the prioritization of inspections and the pursuit of sustainable water resource management.

References

1. United Nations, Sustainable Development Goal 6: Ensure Availability and Sustainable Management of Water and Sanitation for All, URL: https://sdgs.un.org/goals/goal6
2. World Health Organization (WHO), Guidelines for Drinking-Water Quality: Fourth Edition Incorporating the First Addendum (WHO, Geneva, 2017), URL: https://www.who.int/publications/i/item/9789241549950
3. K.G. Liakos, P. Busato, D. Moshou, S. Pearson, D. Bochtis, Machine learning in agriculture: a review, Sensors 18(8), 2674 (2018). DOI: 10.3390/s18082674
4. F. García-Ávila, A. Avilés-Añazco, J. Ordoñez-Jara, C. Guanuchi-Quintero, L. Flores del Pino, L. Ramos-Fernández, Application of machine learning techniques for the prediction of water quality index in an Andean city, Environ. Monit. Assess. 195(1), 198 (2023). DOI: 10.1007/s10661-022-10806-1
5. M.A. Wani, F.A. Bhat, S. Afzal, A.I. Khan, Advances in Deep Learning (Springer, Singapore, 2020). DOI: 10.1007/978-981-13-6794-6
6. T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed. (Springer, New York, 2009), URL: https://hastie.su.domains/ElemStatLearn/

7. L. Breiman, Random forests, Mach. Learn. 45(1), 5–32 (2001). DOI: 10.1023/A:1010933404324
8. Kaggle, Water Quality Dataset, URL: https://www.kaggle.com/datasets/adityakadiwal/water-potability
9. W. McKinney, Data structures for statistical computing in Python, in Proceedings of the 9th Python in Science Conference (SciPy 2010) (2010), pp. 56–61. DOI: 10.25080/Majora-92bf1922-00a
10. G. James, D. Witten, T. Hastie, R. Tibshirani, An Introduction to Statistical Learning: With Applications in R (Springer, New York, 2013)
11. F. Pedregosa, G. Varoquaux, A. Gramfort et al., Scikit-learn: machine learning in Python, J. Mach. Learn. Res. 12, 2825–2830 (2011)
12. M. Ali, PyCaret: An Open Source, Low-Code Machine Learning Library in Python, PyCaret Version 3.0.0 (2023)
13. M. Sokolova, G. Lapalme, A systematic analysis of performance measures for classification tasks, Inf. Process. Manag. 45(4), 427–437 (2009). DOI: 10.1016/j.ipm.2009.03.002
14. A. Singh, A.K. Pandey, Water potability classification using machine learning algorithms, in Proceedings of the International Conference on Innovative Computing and Communications (ICICC), 1–9. (2022)
15. L. Chen, H. Wang, H. Li, A comparative study of machine learning models for predicting water potability, Water 15(5), 987 (2023). DOI: 10.3390/w15050987